Stochastic Proximal Iteration: A Non-Asymptotic Improvement Upon Stochastic Gradient Descent

Ernest K. Ryu Stephen Boyd

Abstract

In many stochastic optimization problems, the learner is provided with random functions, and the common practice has been to differentiate the said functions and perform stochastic gradient descent (SGD). However, to use just the gradient and not the entire function seems suboptimal. To address this issue, we present an algorithm which we call stochastic proximal iteration (SPI). Each iterate of SPI is obtained by applying the proximal mapping with respect to the given random function to the previous iterate. This makes SPI an online algorithm with a computational cost comparable to that of SGD in certain interesting settings. Using machinery from monotone operator theory, we show that SPI has an asymptotic performance equivalent to SGD but does better in the non-asymptotic regime.

Contents

1 Introduction

In many machine learning and statistics problems the learner is given a sequence of i.i.d. *random functions*, $f_1(x), \ldots, f_k(x)$, and is asked to find iterates $x_1, \ldots x_k$ that minimize $F(x) = \mathbf{E}f_k(x)$ approximately.

There are two simple and standard methods: The first is to report $x_k = \arg\min_x \sum_{i=1}^k f_k(x)$, the *empirical risk minimizer*. While this approach may produce good iterates x_k's, the computation of each x_k requires a numerical solve of a convex optimization problem, and this may be prohibitively expensive when k and the dimension of x is large.

The second approach is to compute a subgradient $g \in \partial f_k(x_k)$ (so $g = \nabla f_k(x_k)$ if f_k is differentiable) and perform stochastic gradient descent (SGD). Indeed, this is computationally tractable. However, at each iteration we only use a subgradient, which is just part of the provided information. Could we design a better algorithm by using the entire function f_k at each iteration?

In this paper, we present an algorithm that does this: stochastic proximal iteration (SPI). At each iteration of SPI, we apply the proximal mapping with respect to f_k, the random function, to the current iterate, x_{k-1}, to obtain the next iterate, x_k. In Section 2 we set up notation and define the algorithm precisely.

To analyze SPI, we use tools from monotone operator theory and a weaker notion of strong convexity, which we introduce and discuss in Section 3.

Because SPI is applied to the problem setting where SGD is commonly used (and because it was conceived as an improvement upon SGD), we analyze the algorithm's performance with a focus on the comparison with SGD. It turns out that SPI has essentially the same asymptotic performance as SGD but does much better in the non-asymptotic regime. This is discussed in detail in Section 4 and Section 5, respectively.

However, SPI would be meaningless unless it is computationally tractable because of the computationally expensive but well-performing alternative, the empirical risk minimizer. In Section 6 we discuss the computational cost and show that for many interesting examples SPI is tractable and in fact has computational cost comparable to SGD.

Finally, we conclude the paper by exhibiting examples in Section 7 and providing a summary and possible extentions in Section 8.

Note. Most proofs of the presented results are deferred to the appendix in which sections are labeled by alphabets.

2 Problem Setup and Algorithms

2.1 Notation and Terminology

Throughout this paper, all functions will be closed, proper, and convex, and we will often not bother to state this. We do so because a convex function that is not closed or proper may not have a well-defined proximal mapping and because such functions are mere pathologies that do not arise in practice.

Unless stated otherwise, f denotes a function, x denotes the optimization variable in $\mathbf{R}^n$, and ω denotes a random variable in Ω, a sample space. We write $\partial f(x)$ to denote the subdifferential, the set of subgradients. If f is differentiable at x then $\partial f(x) = \{\nabla f(x)\}$. For the definition relating to convex functions, we refer interested readers to Rockafellar's book [Roc70].

Finally, we define the proximal mapping with respect to a function f with step size α as

$$T_{\alpha f}(x) = \operatorname*{arg\,min}_{y \in \mathbf{R}^n} \left\{ \alpha f(y) + \frac{1}{2}\|y - x\|^2 \right\}.$$

(The arg min exists and is unique [PB13].) We will often drop the subscript and write $Tx = T_{\alpha f}x$ when the meaning is clear from context.

2.2 Random Function Oracle Model

The goal is to minimize the convex function defined as

$$F(x) = \mathbf{E}_\omega f(x; \omega),$$

where $\omega \in \Omega$ is a random variable and $f(x; \omega)$ is convex in x for every ω. We denote an optimal point $x^\star$.

We assume that at the kth iteration we are given a random function $f_k(x)$. Equivalently, we are given a sample ω_k and we can evaluate the random function $f(x; \omega_k) = f_k(x)$.

Under certain regularity conditions that allow us to change the order of differentiation and expectation, we have

$$\nabla F(x) = \mathbf{E}_\omega \left[\nabla_x f(x; \omega)\right] = \mathbf{E}_\omega \left[\nabla_x f_k(x)\right].$$

Therefore we can view $\nabla_x f_k(x_k)$ as a noise corrupted version of $\nabla F(x_k)$. (The same holds for subgradients.)

Finally, we assume $F(x) < \infty$ somewhere, and, to avoid pathologies where $F(x) = -\infty$ for some x, we assume $\mathbf{E}_\omega \inf_x f(x; \omega) > -\infty$.

2.3 Stochastic Gradient Descent

First developed by Robbins and Monro [RM51] and extended to non-differentiable functions by Shor [Sho62], stochastic gradient descent (SGD) is a very widely used algorithm that has the following extremely simple form:

$$x_k = x_{k-1} - \alpha_k g_k. \tag{1}$$

where $\mathbf{E}g_k \in \partial F(x_{k-1})$, *i.e.*, g_k is a noisy subgradient, and $x_0 \in \mathbf{R}^n$ is an arbitrary starting point. In our setting, we can take any $g \in \partial f_k(x_{k-1})$. Standard references of SGD include [NY83, SKR85, Pol87, KY03].

The nonnegative parameter α_k, called the *step size*, represents the extent to which we trust our current iterate; if α_k is small we do not deviate much from x_k as we have much confidence in it and vice versa. There are many possible choices of α_k, but in this paper we only consider two types: $\alpha_k = C/k$ and $\alpha_k = \alpha$.

In a model where one is given noisy subgradients, perhaps SGD is the best one can do. In our setting, however, we are given *random functions*, which contain more information than just subgradients.

An interpretation of SGD via the proximal mapping is

$$x_k = T_{\alpha_k \overline{f}_k}(x_{k-1}), \tag{2}$$

where $\overline{f}_k(x) = f_k(x_k) + g_k^T x$ is a *linear model* of f_k. The equivalence between (1) and (2) is readily verified by taking derivatives.

So SGD is the proximal mapping applied to a linear model of the random function f_k. However, why not use the original function instead?

2.4 Stochastic Proximal Iteration

The formulation of Equation (2) suggests the fix

$$x_k = T_{\alpha_k f_k}(x_{k-1}) = \arg\min_x \left\{ \alpha_k f_k(x) + \frac{1}{2}\|x - x_{k-1}\|_2^2 \right\}.$$

We call this algorithm *stochastic proximal iteration* (SPI). Like the counterpart in SGD, the step size α_k encodes our trust in the current iterate and $x_0 \in \mathbf{R}^n$ is an arbitrary starting point.

SPI has the following interpretation: At the kth iteration, we are given f_k, which measures, say, the discomfort of the kth experience. We would

like to choose the next iterate x_k to accomodate the current experience f_k, but, at the same time, we do not want to deviate too much from our current iterate x_{k-1} which we trust by $1/\alpha_k$. Therefore, we let x_k minimize f_k with a penalty, proportional to $1/\alpha_k$, for deviating from x_{k-1}.

This algorithm is not new. First of all, proximal iteration has been analyzed in the deterministic setting in great detail under the name proximal minimization or proximal-point algorithm [Roc76, BC11]. We will discuss some of these results in Section 3.1.

In the field of online convex optimization, algorithms of this spirit are generically referred to as *implicit updates* and convergence results are proved in certain special cases [KB10, McM11, KL11].

Also, Bertsekas studied this algorithm under the name *incremental proximal algorithms* to minimize a *finite* sum of functions [Ber11]. However, his convergence bounds depend on the number of terms in the finite sum and therefore is limited to the finite sum setting.

So to the best of our knowledge, there have not been any convergence results for SPI that are comparable to that of SGD.

3 Analysis preliminaries

It turns out that standard approaches used to analyze SGD do not easily apply to SPI, and this is probably due to the implicit definition of iterates. So rather, our approach in this work has been to use tools from monotone operator theory. In this section, we introduce some standard concepts from monotone operator theory and then provide and explain some new defitions we use later.

3.1 Proxiaml mapping as a Contraction Maping

In this section we analyze the proximal iteration in the non-stochastic setting since understanding of the deterministic setting is important for understanding the stochastic counterpart.

The key observation is that the proximal iteration is a series of contraction mappings. To begin with, the proximal mapping is always non-expansive [Mor65], *i.e.*,

$$\|T_f x - T_f y\| \leq \|x - y\|.$$

If f is μ-strongly convex then

$$\|T_f x - T_f y\| \leq \frac{1}{1+\mu}\|x - y\|,$$

i.e., T_f is a strict contraction [BMMW12]. Moreover, if $x^\star$ is minimizes f, then $x^\star$ is a fixed-point of T_f, *i.e.*, $T_f x^\star = x^\star$ [PB13].

The strong convexity parameter μ represents the curvature of the function. Intuitively speaking, stronger the curvature, stronger the contraction is, and stronger the contraction, faster the convergence is. As the worst-case, the proximal mapping with respect to a linear function offers no contraction.

Finally, with a standard contraction argument we conclude

$$\begin{aligned}\|x_k - x^\star\| = \|T_f x_{k-1} - T_f x^\star\| &\leq \frac{1}{1+\mu}\|x_{k-1} - x^\star\| \\ &\leq \cdots \leq \frac{1}{(1+\mu)^k}\|x_0 - x^\star\|.\end{aligned}$$

So x_k converges exponentially (linearly) to $x^\star$. The idea is that every application of the proximal mapping shrinks our iterate towards the fixed point, $x^\star$.

This simple analysis, however, doesn't immediately apply to the stochastic setting: Even when the *random* mapping T_{f_k} provides a strict contraction, $x^\star$ is no longer a fixed-point.

3.2 M-Rectricted Strong Convexity

In the deterministic setting, we assumed strong convexity, which, intuitively speaking, requires the function to have curvature in all directions. In the stochastic setting, however, this is too strong of an assumption. As a quick example, a random function of the form $f(\omega_k^T x)$ only varies along the one dimension parallel to ω_k and therefore can have curvature only along that one direction. Therefore, we introduce a weaker notion to replace strong convexity.

A convex function is μ-strongly convex if for any $x, x_0 \in \mathbf{dom} f$ and $g \in \partial f(x_0)$ we have

$$f(x) \geq f(x_0) + g^T(x - x_0) + \frac{\mu}{2}\|x - x_0\|^2.$$

On the other hand, we say f satisfies M-restricted strong convexity if for any $x, x_0 \in \mathbf{dom} f$ and $g \in \partial f(x_0)$

$$f(x) \geq f(x_0) + g^T(x - x_0) + \frac{1}{2}(x - x_0)^T M(x - x_0)$$

holds, where M is a symmetric positive semidefinite matrix. Equivalently, f satisfies M-restricted strong convexity if $f(x) - \frac{1}{2}x^T Mx$ is a convex function (*c.f.* Section B). If $M \succeq \mu I$ then f is μ-strongly convex.

For each given random function f, there is an associated strong convexity matrix M, where M could simply equal the 0 matrix in the worst case. (M is not unique, *c.f.* Section E.)

To prove convergence of SPI, we assume that $\mathbf{E}M \succ 0$, *i.e.*, $\lambda_{\min}(\mathbf{E}M) > 0$, instead of μ-strong convexity. Intuitively, this assumption requires that the random proximal mapping T_f has positive probability to contract in any direction, although not in all directions simultaneously.

Finally, we will note that the assumption $\lambda_{\min}(\mathbf{E}M) > 0$ implies $F(x)$ is $\lambda_{\min}(\mathbf{E}M)$-strongly convex and therefore has a unique minimum, which we denote as $x^\star$ (*c.f.* Section B).

3.3 Contraction in Expectation

For our convergence results, we need the proximal mapping $T_{\alpha f}$ to be a contraction in expectation, *i.e.*,

$$\mathbf{E}\|T_{\alpha f}x - T_{\alpha f}y\|^2 \leq \gamma(\alpha)\|x - y\|^2 \tag{3}$$

for all $x, y \in \mathbf{R}^n$, where $\gamma(\alpha) < 1$ for $\alpha > 0$. We also need to know the behavior of the *contraction factor*, $\gamma(\alpha)$, as $\alpha \to 0$.

However, directly analyzing the contraction factor of $T_{\alpha f}$ is hard. So instead we analyze the contraction factor of a quadratic model of f.

Given f with restricted strong convexity matrix M and a subgradient g at x_0, we define a *quadratic model* as

$$\breve{f}(x) = f(x_0) + g^T(x - x_0) + \frac{1}{2}(x - x_0)^T M(x - x_0).$$

Since f is the sum of two convex functions,

$$f(x) = \breve{f}(x) + (f(x) - \breve{f}(x)),$$

f has "stronger curvature" than $\breve{f}$. (Since $f - \breve{f} = f - \frac{1}{2}x^T Mx + \text{affine}$, it is convex.) Therefore, it seems intuitively reasonable to expect $T_{\alpha f}$ to be a stronger contraction than $T_{\alpha \breve{f}}$. This assertion turns out to be somewhat true.

Theorem 1. *Assume $\breve{f}$ is a random convex quadratic, for which*

$$\mathbf{E}\frac{\|T_{\breve{f}}x - T_{\breve{f}}y\|^2}{\|x-y\|^2} \leq \gamma^2$$

holds for any $x, y \in \mathbf{R}^n$, and r is random closed proper convex function. Then we have

$$\mathbf{E}\frac{\|T_{\breve{f}+r}x - T_{\breve{f}+r}y\|^2}{\|x-y\|^2} \leq \gamma.$$

Note that absense of the square in the second equation; this is saddening since $\gamma^2 \leq \gamma \leq 1$. However, the theorem would not be true with the square (*c.f.* Section E).

Theorem 1, which is the backbone of the theoretical analysis of this paper, tells us that although $T_{\alpha f}$ is not necessarily a stronger contraction than $T_{\alpha \breve{f}}$, we can still bound the contraction factor of $T_{\alpha f}$ with that of $T_{\alpha \breve{f}}$.

With straightforward calculations we get

$$T_{\alpha \breve{f}}(x) - T_{\alpha \breve{f}}(y) = (I + \alpha M)^{-1}(x - y).$$

(Although the function $\breve{f}$ depends on the choice of x_0 and g, this choice ultimately does not matter as the above equation does not depend on it.) This in turn gives us

$$\mathbf{E}\|T_{\alpha \breve{f}}(x) - T_{\alpha \breve{f}}(y)\|^2 = (x-y)^T \mathbf{E}(I + \alpha M)^{-2}(x-y)$$
$$\frac{\mathbf{E}\|T_{\alpha \breve{f}}(x) - T_{\alpha \breve{f}}(y)\|^2}{\|x-y\|^2} \leq \lambda_{\max}\left(\mathbf{E}(I + \alpha M)^{-2}\right).$$

Now we define the function

$$\gamma^2(\alpha) = \lambda_{\max}\left(\mathbf{E}(I + \alpha M)^{-2}\right).$$

Combining this with Theorem 1, we arrive at Equation (3).

Finally, the following theorem characterizes the contraction factor $\gamma(\alpha)$.

Theorem 2. *If $\mathbf{E}M$ is well-defined (i.e., finite-valued) and satisfies $\lambda_{\min}(\mathbf{E}M) > 0$, the contraction factor $\gamma(\alpha)$ (defined for $\alpha \geq 0$) is a strictly decreasing function with*

$$\gamma(\alpha) = 1 - \alpha\lambda_{\min}(\mathbf{E}M) + o(\alpha)$$

as $\alpha \to 0$.

3.4 Quantifying Randomness

Since SPI operates on random data, its convergence will depend on the randomness, or noise level, of the setting. As we will see in Sections 4.3 and 5.4, this dependence is only through $\|Tx^\star - x^\star\|$, a quantity that is zero in the deterministic setting as discussed in Section 3.1 and bounded by the following lemma.

Lemma 1. *Given any convex function f and a point x in $\mathbf{dom} f$, where $\partial f(x) \neq \emptyset$, we have*

$$\|Tx - x\| \leq \|g(x)\|,$$

where g is the minimum-norm subgradient, i.e.,

$$g(x) = \operatorname*{arg\,min}_{v \in \partial f(x)} \|v\|.$$

($g(x)$ is unique, *c.f.* Section B.) Therefore $\mathbf{E}\|T_{\alpha f}x^\star - x^\star\|^2 \leq \alpha^2 \mathbf{E}\|g(x^\star)\|^2$. Since $\mathbf{E}\nabla f(x^\star) = 0$ in the differentiable case, $\mathbf{E}\|g(x^\star)\|^2$ can be thought of as a noise level.

4 Asymptotic Analysis

In this section, we analyze the asymptotic behavior of SPI and compare it to SGD. As we will see, the *asymptotic* performances of the two algorithms are the same (in the big O sense).

4.1 Asymptotic Behavior of the Proximal Mapping

The limiting behavior of $T_{\alpha f}(x)$ as $\alpha \to 0$ is crucial to the asymptotic analysis and intuitively justifies why SPI behaves like SGD, asymptotically. When f is twice continuously differentiable, the asymptotic behavior is

$$T_f(x) = x - \alpha \nabla f(x) + \alpha^2 (\nabla^2 f(x)) \nabla f(x) + o(\alpha^2),$$

which is readily obtained by applying a basic Neumann series argument to Parikh's characterization [PB13].

However, as we wish to deal with non-differentiable convex functions as well, a more general result is necessary. The following result is, to the best of our knowledge, novel.

Theorem 3. *Let f be any closed proper convex function and x a point in* $\mathbf{dom} f$*, where $\partial f(x) \neq \emptyset$. Then*

$$T_f(x) = x - \alpha g(x) + o(\alpha),$$

where g is the minimum-norm subgradient.

The theorem states that, for small α, the proximal mapping is, to first order, a subgradient step with respect to the subgradient of smallest magnitude.

4.2 Asymptotic Rate of Convergence

Because of Theorem 12, we expect the asymptotic performance of SGD and SPI to be equivalent. The following theorem affirms this intuition.

Theorem 4. *Assume that:*

- *the random functions f_k are almost surely twice continuously differentiable at any given point, have a common closed domain, and associated restricted strong convexity matrices M_k,*
- $\mathbf{E}M$ *is well defined,*
- $\lambda_{\min}(\mathbf{E}M) > 0$*,*
- $\|\nabla f(x) - \nabla f(y)\| \leq \widetilde{L}\|x - y\|$ *for all* $x, y \in \mathbf{dom} f$*, where* $\mathbf{E}\widetilde{L}^2 < \infty$*,*
- $\sigma = \mathbf{E}\|g(x^\star)\|^2 < \infty$*,*
- $x^\star$ *is in the relative interior of the domain.*

Using step sizes $\alpha_k = C/k$, we get

$$\mathbf{E}\|x_k - x^\star\|^2 = \begin{cases} \mathcal{O}(1/k) & \text{if } C\lambda_{\min}(\mathbf{E}M) > 1 \\ \mathcal{O}(\log k/k) & \text{if } C\lambda_{\min}(\mathbf{E}M) = 1 \\ \mathcal{O}(1/k^{C\lambda_{\min}(\mathbf{E}M)}) & \text{if } C\lambda_{\min}(\mathbf{E}M) < 1. \end{cases}$$

These asymptotic results are exactly the same as that of SGD [BM11]. One may be concerned that the rate can be arbitrarily slow if C is too small. However, since this exact phenomenon happens with SGD and already has been studied in detail, we shall not [NY78, NJLS09].

4.3 Proof Outline and Remarks

The proof of Theorem 12 (and of Theorem 14) is based on the decomposition $x_k - x^\star = (Tx_{k-1} - Tx^\star) - (x^\star - Tx^\star)$.

$$\begin{aligned}\mathbf{E}\|x_k - x^\star\|^2 &= \mathbf{E}\|Tx_{k-1} - Tx^\star\|^2 + \mathbf{E}\|Tx^\star - x^\star\|^2 \\ &\quad + \mathbf{E}\langle Tx_{k-1} - Tx^\star, Tx^\star - x^\star\rangle \\ &\leq \gamma(\alpha_k)\|x_{k-1} - x^\star\|^2 + \alpha_k^2\sigma^2 \\ &\quad + \mathbf{E}\langle x_{k-1} - x^\star + \mathcal{O}(\alpha_k), \alpha_k\nabla f(x^\star) + \mathcal{O}(\alpha_k^2)\rangle \\ &= (1 - \mathcal{O}(\alpha_k))\|x_{k-1} - x^\star\|^2 + \mathcal{O}(\alpha_k^2)\end{aligned}$$

The first term is contracted as discussed in Section 3.3, the second term is bounded by noise as discussed in Section 3.4, and the third term is analyzed with the results of Section 4.1. Finally, with a recursive application of the above inequality we get Theorem 12.

We conclude this section with several remarks. Theorem 15 does not hold for non-differentiable random functions, and we presume that the optimal rate for the non-differentiable setting is $\mathcal{O}(1/\sqrt{k})$, as it is with other SGD methods [Ber99, DS09, LLZ09]. However, Theorems 1, 2, and 12 hold for non-differentiable functions and we will utilize this generality in Section 5.

5 Non-Asymptotic Analysis

One can argue that SGD is a reckless algorithm as it blindly moves in the direction of the noisy gradient with no regard to how the step is in fact affecting the objective. On the other hand, SPI is more deliberate as each iterate at least never increases the value on the given function f_k.

As we have already discussed, these differences disappear in the asymptotic regime. Rather, the real merit of SPI comes from its non-asymptotic performance.

In this section, we first discuss how SGD is unstable even with all the standard regularity conditions. We then show that SPI is stable under essentially no assumptions at all. Finally, we show that given the appropriate assumptions, SPI exhibits an exponential rate of convergence into a noise dominated region.

5.1 Instability of SGD

SGD is unstable if initial stepsizes are chosen to be too large. Precisely speaking, $\mathbf{E}\|x_k - x^\star\|_2^2$ can grow exponentially until the step sizes become sufficiently small [BM11].

Theorem 5. *[BM11]. Under certain regularity assumptions, the iterates of SGD with step sizes $\alpha_k = C/k$ satisfy*

$$\mathbf{E}\|x_k - x^\star\|^2 \leq \frac{\exp(D_1 C^2)}{k^{\mu C}} D_2 + \mathcal{O}\left(\frac{1}{k}\right)$$

for $C > 2/\mu$, where D_1, D_2 and μ are certain constants.

This bound (although it has the asymptotic rate $\mathcal{O}(1/k)$) grows exponentially for a given n if we use too large of a C.

This is not merely an artifact of their bound. Consider the following counter example: $f(x) = x^2/2$, no noise, constant step-size $\alpha_k = 3$, and starting point $x_0 = 1$. With simple algebra, we can see that $x_k = (-1)^k 2^k$ and that $\|x_k - x^\star\|$ grows exponentially. If we were to use a decreasing step size, the iterates would eventually converge but the gradient step does harm until the step size reduces to below 2. On the other hand, the SPI iterates with $\alpha_k = 3$ are $x_k = 1/4^k$.

5.2 Stability of SPI

In contrast to SGD, whose iterates may exponentially run away from the solution, SPI is much more stable: Even with essentially no assumptions, SPI iterations can do harm no more than the sum of the step sizes.

Theorem 6. *Assume that:*

- *the set of optimal points, $X^\star$, is nonempty,*
- $\sigma = \sup_{x^\star \in X^\star} \mathbf{E}\|g(x^\star)\| < \infty$,

but nothing else. Then

$$\mathbf{E}D(x_k, X^\star) \leq D(x_0, X^\star) + \sigma \sum_{i=1}^{k} \alpha_i,$$

where $D(x, X^\star)$ denotes the distance of x to the set of optimal points, i.e., $D(x, X^\star) = \inf_{x^\star \in X^\star} \|x - x^\star\|_2$.

5.3 Exponential Convergence to Noise Dominated Region

In machine learning literature it is often loosely argued that SGD performs well because it exhibits rapid convergence to the region where noise becomes dominant [SD03]. As discussed in Section 5.1, this is not always true with SGD, but it is with SPI.

Theorem 7. *Assume that:*

- *the random functions f_k have a common closed domain and associated restricted strong convexity matrices M_k,*
- $\lambda_{\min}(\mathbf{E}M) > 0$,
- $\sigma = \mathbf{E}\|g(x^\star)\| < \infty$.

Using constant step size $\alpha_k = \alpha$, we get

$$\mathbf{E}\|x_k - x^\star\| \leq \gamma^{k/2}(\alpha)\|x_0 - x^\star\| + \frac{\alpha\sigma}{1 - \gamma^{1/2}(\alpha)}.$$

While this bound does *not* show $x_k \to x^\star$, it does inform us of SPI's non-asymptotic behavior: With constant step size, x_k converges exponentially into a neighborhood of $x^\star$ with radius $\alpha\sigma/(1 - \gamma^{1/2}(\alpha))$.

5.4 Proof Outline and Remarks

The proof of Theorem 14 follows from the decomposition of Section 4.3 and an application of the triangle inequality.

$$\begin{aligned}\mathbf{E}\|x_k - x^\star\| &\leq \mathbf{E}\|Tx_{k-1} - Tx^\star\| + \mathbf{E}\|Tx^\star - x^\star\| \\ &\leq \gamma^{1/2}(\alpha)\|x_{k-1} - x^\star\| + \alpha\sigma\end{aligned}$$

With a recursive application of the above we get Theorem 14. Theorem 6 is also shown similarly. Figure 1 illustrates the intuition of this inequality.

Note that the results proven with the triangle inequality, Theorem 6 and 14, do not depend on the differentiability of f. However, this approach does not show the convergence of x_k when applied to decreasing step sizes (*c.f.* Section D), which is why in Section 4 we considered the *squared* norm instead.

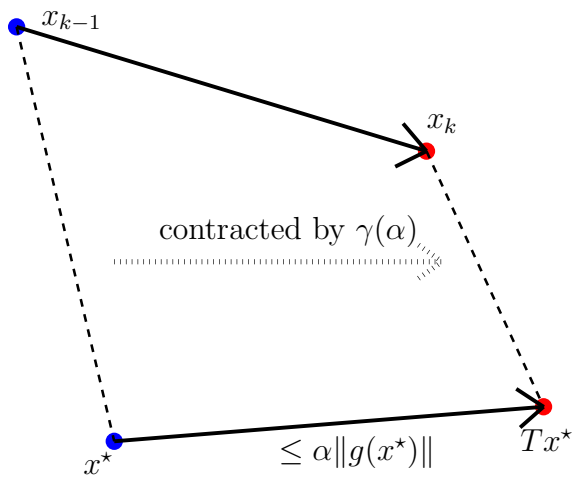

Figure 1: Illustration of the triangle inequality. The distance between $x_k = Tx_{k-1}$ and $Tx^\star$ is contracted but $Tx^\star$ moves away from $x^\star$. Since the contraction is multiplicative, if x_{k-1} is far away from $x^\star$, the next iterate x_k is closer to $x^\star$ than x_{k-1}.

6 Computational Efficiency

One indismissible advantage of SGD is its computational efficiency; each iteration costs $\mathcal{O}(n)$ flops. To be competitive as a practical algorithm, SPI must also have a low computational cost, especially because of the computationally expensive but well-performing alternative $x_k = \arg\min_x \sum_{i=1}^k f_k(x)$, the empirical risk minimizer.

Each iteration of SPI evaluates a proximal mapping, which is a convex optimization problem that in general costs roughly $\mathcal{O}(n^3)$ flops. While it is true that the burden can be alleviated with methods like warm-starting or inexact solves, the $\mathcal{O}(n^3)$ cost is hard to justify. However, in many interesting cases, the iteration can be done with $\mathcal{O}(n)$ cost, if the optimization is done right.

In this section, we will provide several examples where the computational cost of SPI is tractable. Our discussion will be informal and quick, as the ideas we introduce in this section are completely standard in optimization. The purpose is to demonstrate the practical feasibility of SPI in some interesting settings.

6.1 Univariate Function with Seprable Regularization

Consider the class of functions of the form

$$f(\omega^T x) + \sum_{i=1}^{n} r_i(x_i),$$

where x_i here denotes the ith coordinate of x, not the iterate of SPI. For example, MAP with generalized linear models have this form.

First of all, if $r_i = 0$ for all i, then $T_f x_0$ reduces to the univariate optimization problem

$$\text{minimize} \quad f(\omega^T x_0 + a\|\omega\|^2) + \tfrac{a^2}{2}\|\omega\|^2,$$

where $a \in \mathbf{R}$ is the optimization variable. Univariate optimization problems can be solved (say, with bisection) in essentially $\mathcal{O}(n)$ time.

Otherwise, when the r_i's are not zero, we use Lagrange duality for an efficient solution. Consider an equivalent form of the proximal mapping

$$\begin{array}{ll} \text{minimize} & f(y) + \sum_{i=1}^n r_i(x_i) + \frac{1}{2}\|x - x_0\|^2 \\ \text{subject to} & y = \omega^T x, \end{array}$$

where $y \in \mathbf{R}$ and $x \in \mathbf{R}^n$ are the optimization variables. We take the dual of this problem and get

$$\text{maximize} \quad -f^*(-\nu) - \sum_{i=1}^n r_i^\diamond(\nu\omega_i) + \tfrac{\|x_0\|^2}{2},$$

where $\nu \in \mathbf{R}$ is the optimization variable. We write f^* for the Fenchel conjugate of f and

$$r_i^\diamond(\mu) = r_i^*(\mu) \diamond \frac{(\mu + (x_0)_i)^2}{2},$$

where $\diamond$ denotes the infimal convolution [Roc70]. This univariate optimization problem (assuming f^* and $r^\diamond$ are easy to evaluate) can be solved in essentially $\mathcal{O}(n)$ time.

Finally, once the optimal $\nu^\star$ is computed, $x^\star$ can be found as the solution of

$$\text{minimize} \quad \sum_{i=1}^n r_i(x_i) + \tfrac{1}{2}\|x - x_0\|^2 - \nu^\star \omega^T x,$$

where $x \in \mathbf{R}^n$ is the optimization variable. The optimization problem is separable in each coordinate of x and therefore can be solved in essentially $\mathcal{O}(n)$ time. For a more thorough treatment of this idea, interested readers should refer to §5.5.5 of Boyd and Vandenberghe's book [BV04].

6.2 Problems with Structured KKT Systems

When one uses an interior point method to solve an optimization problem, the bottleneck of the computational cost is solving the KKT system. In general, KKT systems of optimization problems cost $\mathcal{O}(n^3)$ time to solve. However, the KKT system sometimes has exploitable structure, which can be used to significantly reduce the cost of the solve. In this section, we provide an example that illustrates this point.

Consider the following function

$$f(\omega^T x) + \|Dx\|_1,$$

where D is a tridiagonal matrix (*e.g.*, a 1D finite difference matrix).

One way to solve the optimization problem of T_f is via a *barrier method* in which we solve a sequence of problems:

$$\begin{array}{ll} \text{minimize} & f(x^T\omega) + \frac{1}{2}\|x - x_0\|^2 + \mathbf{1}^T y \\ & -\frac{1}{t}\sum_{i=1}^n \log(y_i^2 - (D_i x)^2) \end{array}$$

where $x, y \in \mathbf{R}^n$ are the optimization variables and $t > 0$ is a parameter for the barrier [BV04]. The dominant cost for the barrier method is inverting the KKT matrix of the system. Consider the first $n \times n$ block of the Hessian

$$f''(\omega^T x)\omega\omega^T + I + \frac{2}{t} D^T \mathbf{diag}\left(\frac{y_i^2 + (D_i x)^2}{(y_i^2 - (D_i x)^2)^2}\right) D.$$

The tridiagonal matrix D can be inverted in $\mathcal{O}(n)$ time, and so can this entire sub-block with the Sherman-Morrison-Woodbury formula [GL12]. Finally, the entire KKT system can be solved in $\mathcal{O}(n)$ time by applying block elimination and repeating the same idea [BV04].

6.3 Switch-to-SGD Heuristic

Even in the above examples where we argue the computational cost is $\mathcal{O}(n)$, SPI is still more expensive than SGD (at least by a constant factor). Therefore, it would make sense to switch from SPI to SGD once we reach the asymptotic regime.

One heuristic criterion for this is to run SPI until

$$\|T_{\alpha_k f_k} x_{k-1} - x_{k-1} + \alpha_k \nabla f_k(x_{k-1})\| \leq \varepsilon \alpha_k \|\nabla f_k(x_{k-1})\|$$

for N consecutive iterations, where $\varepsilon > 0$ is a predefined parameter. (When f_k is not differentiable, we can replace ∇f_k with g, the minimum-norm subgradient.)

Intuitively speaking, the method switches to SGD once the SGD and SPI steps no longer differ by much. In the experiments of Section 7, we will see that the performance of this heuristic is essentially as good as that of SPI.

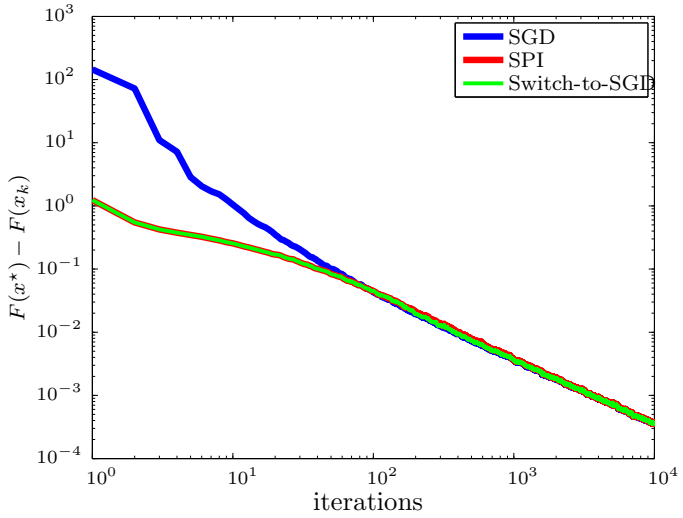

Figure 2: Convergence of the SVM example. SPI and switch-to-SGD heuristic have nearly indistinguishable performance.

7 Examples

In this section, we present examples to compare the performance of SGD, SPI, and the switch-to-SGD heuristic.

The experimental results agree with the theory: SPI indeed performs much better in the non-asymptotic regime but SGD eventually catches up. Also, the switch-to-SGD heuristic performs essentially as good as SPI.

7.1 Support Vector Machine

We first consider the Support Vector Machine (SVM) with hinge loss:

$$\text{minimize} \quad \mathbf{E} \max\{0, 1 - y\omega^T x\} + \tfrac{1}{2}\|x\|^2,$$

where $y = \pm 1$ and $\omega \in \mathbf{R}^n$ are random samples and $x \in \mathbf{R}^n$ is the optimization variable.

At each iteration, we are given a sample (y_k, ω_k) which in turn gives us the random function $\max\{0, 1 - y_i \omega_i^T x\} + \frac{1}{2}\|x\|^2$. This random function is not twice differentiable everywhere but it is, almost surely, at any given x if ω, say, has a continuous distribution.

To evaluate Tx_k, we use the dual problem

$$\begin{array}{ll} \text{maximize} & \nu/y - \frac{1}{2(1+\alpha)}\|\nu\omega + x_0\|^2 + \frac{1}{2}\|x_0\|^2 \\ \text{subject to} & 0 \leq \nu/y \leq \alpha \end{array}$$

in the manner described in Section 6.1. Results are shown in Figure 2.

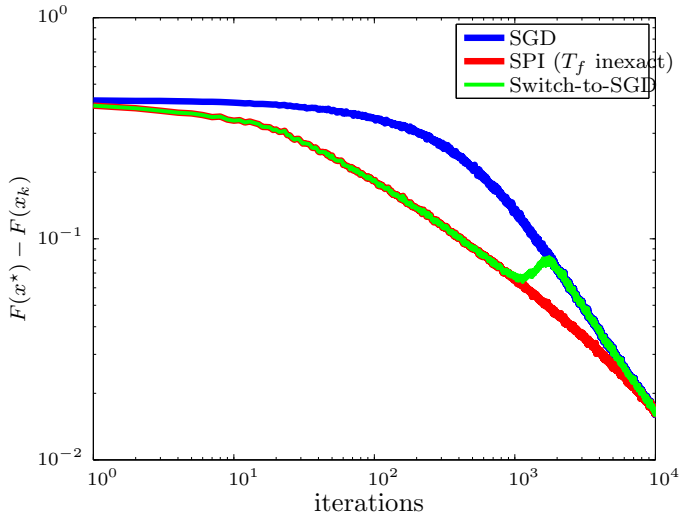

Figure 3: Convergence of the portfolio optimization example. We can see when the switch to SGD happens.

7.2 Portfolio Optimization

Consider the portfolio optimization problem where we have a concave utility function log and the n assets have random returns $\omega > 0$. To maximize the expected utility, we solve

$$\begin{array}{ll} \text{maximize} & \mathbf{E}\log(\omega^T x) \\ \text{subject to} & x \geq 0 \quad \mathbf{1}^T x = 1, \end{array}$$

where $x \in \mathbf{R}^n$ is the optimization variable.

In this example, we approximately evaluate the proximal mapping with the following heuristic. We first evaluate the prox without the constraint $x \geq 0$ (with an analytic formula), and then project this point onto the constraint set $\{x : x \geq 0, \mathbf{1}^T x = 1\}$.

It turns out that this inexact solve approach works quite well. Results are shown in Figure 3.

8 Conclusion and Further Directions

SPI is an online algorithm, applicable to many problems where SGD is used, that has an edge in the non-asymptotic regime. The behavior of SPI is theoretically analyzed and empirically confirmed.

We mention here a few interesting extensions and variations of SPI. The first to relax the strong convexity and differentiability requirement. There has been much work on (the degradation of) SGD's convergence rate without certain assumptions. To study the performance of SPI in these settings would be interesting.

Another direction is to allow inexact solves. Throughout this paper, we have assumed that the proximal mappings are evaluated exactly. By allowing inexact solves, we can reduce the cost of each iteration while perhaps retaining some convergence properties.

Mini-batch updates, where the proximal mapping is with respect to a *sum* of a few of the random functions, is also an interesting topic. The theoretical analysis of this paper does not exclude mini-batch updates and it is a straightforward exercise to understand the dependence of Theorem 4.1 to the batch size. However, the non-asymptotic effect of mini-batching is a bit subtle.

Finally, we can consider the effect of averaging. SGD is often improved with the use of Polyak-Ruppert averaging, $\overline{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_i$ [Rup88, PJ92]. A related idea from monotone operator theory is to use the Cayley operator, which would amount to using the iterates $x_k = 2Tx_{k-1} - x_{k-1}$.

Appendices

The appendix supplements the main document with proofs and details. In section B, we show several results to assure that the objects we use are well-defined. Although these are necessary for the rigorous treatment of the subject, the proofs simply exercise standard techniques and are not particularly intersting. In section C, we provide several lemmas that are used in later proofs but are not directly relevant to core idea of this work. Section D contains the substance of this article, the proofs of the main theorems. Finally, in section E, we provide several counter examples that further illuminate certain concepts.

A Introduction and notation

This article supplements the main document with proofs and details. The material is organized so that the logic is sound with a linear reading, but the material only becomes interesting and important by section D. Therefore, for the casual readers, we suggest quickly reading the statements but not the proofs of section B, skipping section C, and spending time on section D and E.

The numbering of the theorems here is different from that of the main article. (The wording is also sometimes different.) Theorem 1 of the main article corresponds to theorem 10 here, theorem 2 to theorem 11, lemma 1 to lemma 9, theorem 3 to theorem 12, theorem 4 to theorem 15, theorem 6 to theorem 13, and theorem 7 to theorem 14.

Like the main article, we are primarily interested in closed proper convex functions. However, as some of the results are more general, we will not automatically assume these properties. In this article, functions are closed, proper, or convex only when explicitly specified.

Finally, we define some notation. We write $\partial f(x)$ for the set of subdifferentials, $\mathbf{dom} f$ for the domain of the function f, *i.e.*, $\mathbf{dom} f = \{x \in \mathbf{R}^n : f(x) < \infty\}$, and $\mathbf{ri}\, S$ for the relative interior of the set S. A function is closed if its epigraph, $\mathbf{epi}\, f = \{(x, u) \in \mathbf{R}^{n+1} : u \geq f(x)\}$, is closed and is proper if $f = -\infty$ nowhere and $f < \infty$ somewhere. An extreme point x of a convex set S is a point that is an average of two distinct points in S, *i.e.*, $x = (y + z)/2$ for $y, z \in S$ and $y \neq z$. These definitions are standard in convex analysis and can be found in Rockafellar's book [Roc70].

We use $\mathbf{E}$ to denote the expectation with respect to the random variable ω and we write $\mathbf{E}_\omega$ when we wish to emphasize that ω contains the randomness. The norm $\|\cdot\|$ denotes the standard Euclidean norm. The big O notation $f(x) = \mathcal{O}(g(x))$ denotes

$$\lim_{x\to X} \frac{f(x)}{g(x)} = C \notin \{0, \infty, -\infty\},$$

and the small o notation $f(x) = o(g(x))$ denotes

$$\lim_{x\to X} \frac{f(x)}{g(x)} = 0,$$

where the limit X (often ∞) depends on the context. The notation $f(x) \sim g(x)$ and $f(x) \lesssim g(x)$ denotes

$$\lim_{x\to X} \frac{f(x)}{g(x)} = 1 \qquad \lim_{x\to X} \frac{f(x)}{g(x)} \leq 1,$$

respectively, where, again, the limit X (often ∞) depends on the context.

B Sanity check results

Lemma 2 (Uniqueness of minimum-norm subgradient)**.** *If f is subdifferentiable at x then $g(x)$, the minimum-norm subgradient at x, exists and is unique.*

Proof. The set $\partial f(x)$ is closed and convex [Roc70, §23] and is nonempty by assumption. The minimum-norm subgradient is the solution to the optimization problem

$$\begin{array}{ll} \text{minimize} & \|g\|^2 \\ \text{subject to} & g \in \partial f(x), \end{array}$$

where g is the optimization variable. Existence of the solution is readily shown by observing that the sublevel sets of $\|\cdot\|^2$ is compact and applying a simple convergent subsequence argument. Uniqueness of the solution follows from strict convexity of the objective. □

Lemma 3 (Restricted strong convexity)**.** *The two definitions of M-restricted strong convexity are equivalent if f is closed, proper, and convex. In other words, for any closed proper convex function f,*

$$f(x) \geq f(x_0) + g^T(x - x_0) + \frac{1}{2}(x - x_0)^T M(x - x_0) \tag{4}$$

for any $x, x_0 \in \mathbf{dom} f$ and $g \in \partial f(x_0)$ holds if and only if

$$f(x) - \frac{1}{2}x^T Mx$$

is convex.

Proof. We first show the forward direction. Let $x, y \in \mathbf{dom} f$ and $z = \lambda x + (1 - \lambda)y$, where $\lambda \in [0, 1]$.

If $z \in \mathbf{ri}\,(\mathbf{dom} f)$, then a subgradient $g_z \in \partial f(z)$ exists [Roc70, §23]. So we apply Equation (4) to the pairs (y, z) and (x, z) to get

$$f(x) \geq f(z) + g_z^T(x - z) + \frac{1}{2}(x - z)^T M(x - z)$$
$$f(y) \geq f(z) + g_z^T(y - z) + \frac{1}{2}(y - z)^T M(y - z).$$

Now we multiply the first inequality by $(1 - \lambda)$ and the second by λ and add the two to get

$$\lambda\left(f(x) - \frac{1}{2}x^T Mx\right) + (1 - \lambda)\left(f(y) - \frac{1}{2}y^T My\right) \geq f(z) - \frac{1}{2}z^T Mz. \tag{5}$$

Now we extend the result to $z \in \mathbf{dom} f$. Let x_k be a sequence in $\mathbf{ri}\,(\mathbf{dom} f)$ such that $x_k \to x$ and $x, x_1, x_2, \ldots$ all lie on a single line segment. Define y_k in the same manner and write $z_k = \lambda x_k + (1 - \lambda)y_k$. Then we have

$$\lambda\left(f(x_k) - \frac{1}{2}x_k^T Mx_k\right) + (1 - \lambda)\left(f(y_k) - \frac{1}{2}y_k^T My_k\right) \geq f(z_k) - \frac{1}{2}z_k^T Mz_k$$

Since a closed convex function restricted to a line segment is continuous even at the end-points [Roc70, §10] (if the line segment is within $\mathbf{dom} f$) we can take the limit on both sides to get inequality (5) for the general case.

Finally, we show the other direction. Since $f(x) - \frac{1}{2}x^T Mx$ is convex for any $x, x_0 \in \mathbf{dom} f$ and $g \in \partial f(x_0)$ we have

$$f(x) - \frac{1}{2}x^T Mx \geq f(x_0) - \frac{1}{2}x_0^T Mx_0 + g^T(x - x_0) - x_0^T M(x - x_0)$$

Rearranging this gives us Equation (4). □

Lemma 4 (Measurability)**.** *Assume that:*

- *the random functions $f(x;\omega) : \mathbf{R}^n \times \Omega \to (-\infty, \infty]$ are (almost surely) closed, proper, and convex,*
- *$f(x;\omega)$ have a common domain $D \subseteq \mathbf{R}^n$, i.e., $f(x;\omega) < \infty$ for any $x \in D$ (almost surely),*
- *$f(x;\omega)$ are measurable for any given x, with respect to ω's σ-algebra,*
- *ω's σ-algebra is complete.*

Then $T_f x$ and $\inf_{x\in D} f(x;\omega)$ are measurable with respect to ω's σ-algebra.

Proof. We first show the measurability of $\inf_{x\in D} f(x;\omega)$. Since $f(x;\omega)$ is closed, it is lower semi-continuous [Roc70, §7] and continuous in $\mathbf{ri}\,(D)$[Roc70, §10]. Now let Q be a countable dense subset of $\mathbf{ri}\, D$. Then we have

$$\inf_{x\in D} f(x;\omega) = \inf_{x\in \mathbf{ri}\, D} f(x;\omega) \tag{6}$$

$$= \inf_{x\in Q} f(x;\omega), \tag{7}$$

where all equality holds almost surely.

In equation (6), $\inf_{x\in D} f(x;\omega) \leq \inf_{x\in \mathbf{ri}\, D} f(x;\omega)$ is clear since $\mathbf{ri}\, D \subseteq D$. Now consider any sequence x_k in $\mathbf{ri}\, D$ and any point $x \in D$ such that $x_k \to x$ and $x, x_1, x_2, \ldots$ all lie on a single line segment. Since a closed convex function restricted to a line segment is continuous even at the end points [Roc70, §10], we have $f(x;\omega) = \lim_{k\to\infty} f(x_k;\omega) \geq \inf_{x\in \mathbf{ri}\, D} f(x;\omega)$. Therefore $\inf_{x\in D} f(x;\omega) \geq \inf_{x\in \mathbf{ri}\, D} f(x;\omega)$, and we conclude equation (6).

Equation (7) follows from continuity. Since a countable infimum of random variables is measurable [Bil95, §13], equation (7) is measurable.

Finally, $\inf_{x\in D} f(x;\omega)$ is measurable as it differs from a measurable random variable only on a null set and since our probability measure is complete.

Next, we show measurability of $T_f x$. With a some careful thought we get

$$\{\omega \in \Omega : (T_f x(\omega))_i < \alpha\} = \bigcup_{k=1}^{\infty} \bigcap_{y \in Q \cap \{y_i \geq \alpha\}} \left\{\omega \in \Omega : f(y;\omega) + \frac{1}{2}\|y - x\|^2 > m(\omega) + \frac{1}{2^k}\right\}$$

for $i = 1, 2, \ldots, n$ and any $\alpha \in \mathbf{R}$. In other words, the preimage of the coordinatewise sublevel sets are measurable. Since the coordinatewise sublevel sets generate the Borel σ-algebra, this implies that $T_f x(\omega)$ is measurable. □

Lemma 5 (Uniqueness of $x^\star$)**.** *Assume the random convex functions $f(x;\omega)$ are closed and $\mathbf{E}_\omega \inf_x f(x;\omega) > -\infty$. Then $F(x) = \mathbf{E}_\omega f(x;\omega)$ is a closed function. So if $F(x) < \infty$ for some x, then F is closed, proper, and convex. Furthermore, assume each function $f(x;\omega)$ has an associated strong convexity matrix $M(\omega)$ such that $\mathbf{E}M$ is well-defined and $\lambda_{\min}(\mathbf{E}M) > 0$. Then F satisfies $\mathbf{E}M$-restricted strong convexity, and the minimum of $F(x)$ exists and is unique.*

Proof. First assume $\mathbf{E}_\omega \inf_x f(x;\omega) \neq \infty$ because otherwise $F(x) = \infty$ everywhere and there is nothing to show.

Note that F is closed if and only if F is lower semi-continuous [Roc70, §7]. We apply Fatou's lemma [Bil95, §16] to the nonnegative functions $f(x;\omega) - \inf_x f(x;\omega)$ and get

$$\liminf_{x \to x_0} F(x) = \liminf_{x \to x_0} \mathbf{E} f(x;\omega) \geq \mathbf{E} \liminf_{x \to x_0} f(x;\omega) \geq \mathbf{E} f(x_0;\omega) = F(x_0).$$

So F is lower semi-continuous and therefore closed.

Next, we have $F(x) \geq \mathbf{E}_\omega \inf_x f(x;\omega) > -\infty$, so if $F(x) < \infty$ for some x, by definition F is proper.

By lemma 3, $F(x)$ satisfies $\mathbf{E}M$-restricted strong convexity if and only if $F(x) - \frac{1}{2}x^T \mathbf{E}M x$ is convex. Since we have

$$F(x) - \frac{1}{2}x^T \mathbf{E}M x = \mathbf{E}_\omega \left[f(x;\omega) - \frac{1}{2}x^T M(\omega) x \right],$$

where $f(x;\omega) - \frac{1}{2}x^T M x$ is convex by assumption, the expected function is convex. Thus $F(x)$ satisfies $\mathbf{E}M$-restricted strong convexity.

Let x_0 be any point in $\mathbf{ri}\,(\mathbf{dom} F)$ (which exists by [Roc70, §6]). Then

$$F(x) \geq F(x_0) + g^T x + \frac{1}{2}(x - x_0)^T \mathbf{E}M(x - x_0),$$

where $g \in \partial F(x_0)$ (which exists since $x \in \mathbf{ri}\,(\mathbf{dom} F)$ [Roc70, §23]). So any x such that $F(x) \leq F(x_0)$ must be in the compact set

$$K = \left\{ x \in \mathbf{R} : F(x_0) + g^T x + \frac{1}{2}(x - x_0)^T \mathbf{E} M (x - x_0) \leq F(x_0) \right\}.$$

(K is compact since $\mathbf{E}M \succ 0$.) Now let x_k be a sequence such that that $F(x_k) \to \inf_x F(x)$. Then for large enough k, we have

$$(x_k, F_k) \in (\mathbf{epi}\, F) \cap \left(K \times [\inf_x F(x), F(x_0)] \right).$$

Since F is closed, $\mathbf{epi}\, F$ is by definition closed, and since $K \times [\inf_x F(x), F(x_0)]$ is a compact set, (x_k, F_k) has a convergent subsequence that converges to $(x^\star, F(x^\star)) \in \mathbf{epi}\, F$. By construction, $F(x^\star) = \inf_x F(x)$ holds and therefore $x^\star$ is a solution for $F(x)$. So a minimizer of F exists.

Finally, since F is strongly convex, it is strictly convex and therefore the solution is unique. □

Lemma 6 (Swapping expectation and differentiation)**.** *Assume that the random function f is almost surely twice continuously differentiable at x, that $\|\nabla f(x) - \nabla f(y)\| \leq \widetilde{L}\|x - y\|$ for all $x, y \in \mathbf{dom} f$, where $\mathbf{E}\widetilde{L}^2 < \infty$, and that $\mathbf{E}\|g(x^\star)\|^2 < \infty$. (These are assumed in theorem 15.) Then $F(x)$ is differentiable at $x^\star$ and*

$$\nabla F(x) = \mathbf{E}_\omega \nabla_x f(x; \omega).$$

Proof. By definition of differentiation, we have

$$\frac{\partial}{\partial x_i} F(x) = \lim_{h \to 0} \mathbf{E}\left[\frac{f(x + he_i) - f(x)}{h} \right] = \lim_{h \to 0} \mathbf{E}\left[\frac{1}{h} \int_0^h e_i^T \nabla f(x + se_i)\, ds \right].$$

To use dominated convergence, bound the inner term as follows:

$$\begin{aligned} \left| \frac{1}{h} \int_0^h e_i^T \nabla f(x + se_i)\, ds \right| &\leq \frac{1}{h} \int_0^h \|\nabla f(x + se_i)\|\, ds \\ &\leq \sup_{|s| \leq h} \|\nabla f(x + se_i)\| \\ &\leq \|\nabla f(x^\star)\| + \widetilde{L}(\|x^\star - x\| + |h|). \end{aligned}$$

The final term has finite expectation by assumption. Therefore, by dominated convergence, we can swap the order of the limit and expectation, and this proves the desired result. □

C Auxiliary results

Lemma 7. *Let H be a symmetric positive definite matrix such that $\lambda_{max}(H) \leq 1$. Then for any x such that $\|x\|_2 = 1$ we have*

$$\|Hx\|_2^2 \leq x^T H x.$$

Proof. Let $v_1, v_2, \ldots, v_n$ be the orthonormal eigenvectors of H corresponding to their eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$. Then x has the eigenvector expansion

$$x = \alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n.$$

We conclude

$$\|Hx\|_2^2 = \sum_{i=1}^n \alpha_i^2 \lambda_i^2 \leq \sum_{i=1}^n \alpha_i^2 \lambda_i = x^T H x.$$

□

Lemma 8 (Semidefinite dominated convergence.)**.** *Let $X_k \to X$ be a sequence of positive semidefinite matrices such that $0 \preceq X_k \preceq Y$, where $\mathbf{E}Y$ is well-defined and finite. Then*

$$\lim_{k\to\infty} \mathbf{E}X_k = \mathbf{E} \lim_{k\to\infty} X_k = \mathbf{E}X.$$

Proof. First, let $\|\cdot\|_{\max}$ be the entrywise maximum norm and $\|\cdot\|_2$ be the spectral norm. Since all finite dimensional norms are equivalent, there is a $c > 0$ such that

$$\frac{1}{c}\|\cdot\|_{\max} \leq \|\cdot\|_2 \leq c\|\cdot\|_{\max}.$$

Since $0 \preceq X_k \preceq Y$ implies

$$\|X_k\|_2 = \lambda_{\max}(X_k) \leq \lambda_{\max}(Y) = \|Y\|_2,$$

we have

$$|(X_k)_{ij}| \leq \|X_k\|_{\max} \leq c\|X_k\|_2 \leq c\|Y\|_2 \leq c^2\|Y\|_{\max} \leq c^2 \sum_{ij} |Y_{ij}|$$

We conclude by applying dominated convergence theorem to X_k, elementwise. □

Theorem 8 (7.7.4 of Horn and Johnson [HJ90])**.** *If positive semidefinite matrices satisfy $A \succeq B \succeq 0$, then $B^{-1} \succeq A^{-1}$.*

Theorem 9 (24.4 of Rockafellar [Roc70])**.** *Let f be a closed proper convex function. If $x_1, x_2, \ldots$ and $v_1, v_2, \ldots$ are sequences such that $v_i \in \partial f(x_i)$, where x_i converges to x and v_i converges to v, then $v \in \partial f(x)$.*

D Main results

Theorem 10 (Bounding the contraction factor)**.** *Assume $\breve{f}$ is a convex quadratic and r is a closed proper convex function. Then*

$$\frac{\|T_{\breve{f}+r}x - T_{\breve{f}+r}y\|^2}{\|x-y\|^2} \leq \frac{\|T_{\breve{f}}x - T_{\breve{f}}y\|}{\|x-y\|}$$

holds for any $x, y \in \mathbf{R}^n$. Therefore, if $\breve{f}$ and r are random and

$$\mathbf{E}\frac{\|T_{\breve{f}}x - T_{\breve{f}}y\|^2}{\|x-y\|^2} \leq \gamma^2$$

holds for any $x, y \in \mathbf{R}^n$, when we have

$$\mathbf{E}\frac{\|T_{\breve{f}+r}x - T_{\breve{f}+r}y\|^2}{\|x-y\|^2} \leq \gamma.$$

Proof. Write $f = \breve{f} + r$. Assume $x \neq y$ and $T_f x \neq T_f y$ as otherwise there is nothing to show.

Define

$$\tilde{r}_\varepsilon(z) = -(\nabla\breve{f}(T_f x) + T_f x - x)^T z + \frac{1}{2}(z - T_f x)^T \frac{v_\varepsilon v_\varepsilon^T}{a_\varepsilon}(z - T_f x),$$

where

$$\begin{aligned} v_\varepsilon &= (\nabla\breve{f}(T_f x) + T_f x - x) - (\nabla\breve{f}(T_f y) + T_f y - y) + \varepsilon(T_f y - T_f x) \\ a_\varepsilon &= v_\varepsilon^T(T_f y - T_f x) \end{aligned}$$

for some $\varepsilon > 0$. Since we have

$$v_\varepsilon \in -\partial r(T_f x) + \partial r(T_f y) + \varepsilon(T_f y - T_f x)$$

and since ∂r is a monotone operator, we have

$$a_\varepsilon = v_\varepsilon^T(T_f y - T_f x) \geq \varepsilon\|T_f y - T_f x\|^2 > 0,$$

and therefore $\tilde{r}_\varepsilon$ is a well-defined convex quadratic.

By design, $\tilde{r}_\varepsilon$ satisfies

$$\begin{aligned}
\nabla\tilde{r}_\varepsilon(T_f(x)) &= -(\nabla\breve{f}(T_f x) + T_f x - x) \\
\nabla\tilde{r}_\varepsilon(T_f(y)) &= -(\nabla\breve{f}(T_f y) + T_f y - y) + \varepsilon(T_f y - T_f x) \\
T_{\breve{f}+\tilde{r}_\varepsilon} x &= T_{\breve{f}+r} x = T_f x \\
\lim_{\varepsilon\to 0} T_{\breve{f}+\tilde{r}_\varepsilon} y &= T_{\breve{f}+r} y = T_f y. \qquad (8)
\end{aligned}$$

Let $\hat{n} = (x-y)/\|x-y\|$. Now we have

$$\begin{aligned}
\frac{\|T_{\breve{f}}x - T_{\breve{f}}y\|}{\|x-y\|} &= \|(1+M)^{-1}\hat{n}\| && (9)\\
&\geq \hat{n}(1+M)^{-1}\hat{n} && (10)\\
&\geq \hat{n}\left(1 + M + \frac{v_\varepsilon v_\varepsilon^T}{a_\varepsilon}\right)^{-1}\hat{n} && (11)\\
&\geq \left\|\left(1 + M + \frac{v_\varepsilon v_\varepsilon^T}{a_\varepsilon}\right)^{-1}\hat{n}\right\|^2 && (12)\\
&= \frac{\|T_f x - T_{\breve{f}+\tilde{r}_\varepsilon} y\|^2}{\|x-y\|^2} && (13)\\
&\to \frac{\|T_f x - T_f y\|^2}{\|x-y\|^2} \quad \text{as } \varepsilon \to 0. && (14)
\end{aligned}$$

Equation (9) follows from explicitly working out $T_{\breve{f}}x - T_{\breve{f}}y$, (10) from Cauchy-Schwartz, (11) from theorem 8, (12) from lemma 7, (13) again from explicitly working out $T_{\breve{f}+\tilde{r}_\varepsilon}x - T_{\breve{f}+\tilde{r}_\varepsilon}y$, and (14) from equation (8).

Finally, we take expectations on both sides and apply Jensen's inequality [Bil95, §5] to get the desired result.

$$\mathbf{E}\frac{\|T_f x - T_f y\|^2}{\|x-y\|^2} \leq \mathbf{E}\frac{\|T_{\breve{f}}x - T_{\breve{f}}y\|}{\|x-y\|} \leq \left(\mathbf{E}\frac{\|T_{\breve{f}}x - T_{\breve{f}}y\|^2}{\|x-y\|^2}\right)^{1/2} \leq \gamma.$$

□

Corollary 1. *By theorem 10 and the reasoning of section 3.3 of the main article, we conclude that*

$$\mathbf{E}\|T_{\alpha f}x - T_{\alpha f}y\|^2 \leq \gamma(\alpha)\|x-y\|^2,$$

where the asymptotic behavior of γ is characterized by theorem 11. Also, by Jensen's inequality [Bil95, §5], we have

$$\mathbf{E}\|T_{\alpha f}x - T_{\alpha f}y\| \leq \gamma^{1/2}(\alpha)\|x - y\|.$$

Theorem 11 (The contraction factor)**.** *Let $\gamma(\alpha)$ defined as in section 3.3 of the main article. If $\mathbf{E}M$ is well-defined (i.e., finite-valued) and satisfies $\lambda_{\min}(\mathbf{E}M) > 0$, the contraction factor $\gamma(\alpha)$ (defined for $\alpha \geq 0$) is a strictly decreasing function with*

$$\gamma(\alpha) = 1 - \alpha\lambda_{\min}(\mathbf{E}M) + o(\alpha)$$

as $\alpha \to 0$.

Proof. Note that the matrices I, M, $I + \alpha M$, $(I + \alpha M)^{-1}$, and all other variants share a same set of eigenvectors and therefore commute.

We first show that $\gamma(\alpha)$ is a strictly decreasing function. Since $\lambda_{\min}(\mathbf{E}M) > 0$, the event

$$A_{\varepsilon,x} = \{\omega \in \Omega : \varepsilon \leq x^T M x, \},$$

satisfies $\mathbf{P}(A_{\varepsilon,x}) > 0$ for sufficiently small $\varepsilon > 0$. With routine calculations, we find that

$$x^T(I + \alpha M)^{-2}x > x^T(I + (\alpha + h)M)^{-2}x \qquad \text{conditioned on } A_{\varepsilon,x}$$

holds. This implies

$$x^T\mathbf{E}(I + \alpha M)^{-2}x > x^T\mathbf{E}(I + (\alpha + h)M)^{-2}x,$$

which in turn implies

$$\min_{\|x\|=1} \left\{x^T\mathbf{E}(I + \alpha M)^{-2}x - x^T\mathbf{E}(I + (\alpha + h)M)^{-2}x\right\} = \delta > 0,$$

since $\|x\| = 1$ is a compact set. Therefore

$$\mathbf{E}(I + \alpha M)^{-2} \geq \mathbf{E}(I + (\alpha + h)M)^{-2}x + \delta I,$$

and we conclude

$$\gamma^2(\alpha) \geq \gamma^2(\alpha + h) + \delta > \gamma^2(\alpha + h),$$

i.e., $\gamma(\alpha)$ is strictly decreasing.

We now show the asymptotic expansion. First, the following shows $(I + \alpha M)^{-2} = I - 2\alpha M + o(\alpha)$

$$-\lim_{\alpha \to 0^+} \frac{1}{\alpha} \left((I + \alpha M)^{-2} - I \right) = \lim_{\alpha \to 0^+} (I + \alpha M)^{-2} 2M \left(I + (\alpha/2) M \right) = 2M.$$

Next, to change the order of expectation and limit, we use the dominated convergence theorem. The limiting term satisfies

$$\begin{aligned} 0 &\preceq (I + \alpha M)^{-2} 2M \left(I + (\alpha/2) M \right) \\ &\preceq (I + \alpha M)^{-2} 2M \left(I + \alpha M \right) \\ &= (I + \alpha M)^{-1} 2M \preceq 2M \end{aligned}$$

for $\alpha \geq 0$. Since $\mathbf{E}M$ is by assumption well-defined and finite, we use lemma 8, a version of the dominated convergence theorem, to conclude

$$\lim_{\alpha \to 0^+} \frac{1}{\alpha} \left(\mathbf{E}(I + \alpha M)^{-2} - I \right) = 2\mathbf{E}M,$$

which says

$$\mathbf{E}(I + \alpha M)^{-2} = I - 2\alpha \mathbf{E}M + o(\alpha).$$

This, in turn, implies

$$\begin{aligned} \gamma(\alpha) &= \sqrt{\lambda_{\max} \left(I - 2\alpha \mathbf{E}M + o(\alpha) \right)} \\ &= \sqrt{1 - 2\alpha \lambda_{\max} \left(\mathbf{E}M + o(1) \right)} \\ &= 1 - \alpha \lambda_{\max}(\mathbf{E}M) + o(\alpha), \end{aligned}$$

where the last line is by the continuity of $\lambda_{\max}$.

□

Lemma 9 (Bound on the proximal step)**.** *Assime f is subdifferentiable at x. Than for any $g \in \partial f(x)$, we have $\|T_f(x) - x\| \leq \|g\|$. In particular,*

$$\|T_f(x) - x\| \leq \min_{g \in \partial f(x)} \|g\|.$$

Proof. By definition of $T_f x$ we have

$$-(T_f x - x) \in \partial f(T_f x).$$

Since convexity of $f(x)$ implies $\langle \partial f(x_1) - \partial f(x_2), x_1 - x_2 \rangle \geq 0$, we conclude

$$\begin{aligned} \langle -(T_f x - x) - g, T_f x - x \rangle &\geq 0 \\ \|T_f x - x\|^2 &\leq \langle -g, T_f x - x \rangle \\ \|T_f x - x\| &\leq \|g\|, \end{aligned}$$

where the last line is by Cauchy-Schwartz. □

Lemma 10 (Bound on the proximal step 2)**.** *For any $g \in \partial f(x)$ and $g' \in \partial f(x - g)$*

$$\|T_f x - x + g\| \leq \|g - g\|.$$

Proof. By simply rearranging terms we get

$$\begin{aligned} T_f(x) &= \arg\min_y \left\{ f(y) + \frac{1}{2}\|y - x\|^2 \right\} \\ &= \arg\min_y \left\{ f(y) - g^T y + \frac{1}{2}\|y - (x - g)\|^2 \right\} \\ &= T_{f(y) - g^T y}(x - g). \end{aligned}$$

Therefore

$$\|T_f(x) - x + g\| = \|T_{f(y)-g^T y}(x - g) - x + g\| \leq \|g' - g\|,$$

where we have used lemma 9 in the last inequality. □

Theorem 12 (Asymptotic behavior of the proximal mapping)**.** *For any closed proper convex function f and an $x \in \mathbf{dom} f$ such that $\partial f(x) \neq \emptyset$, we have*

$$T_{\alpha f} x = x - \alpha g(x) + o(\alpha)$$

as $\alpha \to 0$, where g is the minimum-norm subgradient at x.

Proof. By definition of the proximal mapping, we have

$$-\frac{T_{\alpha f} x - x}{\alpha} = v_\alpha \in \partial f(T_{\alpha f} x).$$

Moreover, by lemma 9, we have

$$\|v_\alpha\| \leq \|g(x)\|.$$

So $\|T_{\alpha f}x - x\| \leq \alpha\|g(x)\|$ and therefore $T_{\alpha f}x \to x$ as $\alpha \to 0$. Also, the sequence v_α is contained in a compact set and therefore there is a convergent subsequence $v_{\alpha_k} \to v$, where $\alpha_k \to 0$. So by theorem 9 we have $v \in \partial f(x)$. Since we know $\|v\| \leq \|g\|$ and since g is, by definition, the only subgradient that satisfies this, we conclude $v = g$.

Finally, since this argument applies to any convergence subsequence of v_α, we conclude that $\lim_{\alpha\to 0^+} v_\alpha = g$. □

Theorem 13 (SPI's worst case)**.** *Assume that:*

- *the set of optimal points, $X^\star$, is nonempty,*
- $\sigma = \sup_{x^\star \in X^\star} \mathbf{E}\|g(x^\star)\| < \infty$,

but nothing else. Then

$$\mathbf{E}D(x_k, X^\star) \leq D(x_0, X^\star) + \sigma \sum_{i=1}^{k} \alpha_i,$$

where $D(x, X^\star)$ denotes the distance of x to the set of optimal points, i.e., $D(x, X^\star) = \inf_{x^\star \in X^\star} \|x - x^\star\|_2$.

Proof. Let $x^\star$ be any optimal point. Then we have

$$\begin{aligned}
\mathbf{E}\left[D(x_k, X^\star)\right] &\leq \mathbf{E}\|x_k - x^\star\| \\
&\leq \mathbf{E}\|T_{\alpha_k f_k} x_{k-1} - T_{\alpha_k f_k} x^\star\| + \mathbf{E}\|T_{\alpha_k f_k} x^\star - x^\star\| \\
&\leq \mathbf{E}\|x_{k-1} - x^\star\| + \alpha_k \sigma \\
\cdots &\leq \|x_0 - x^\star\| + \sigma \sum_{i=1}^{k} \alpha_i.
\end{aligned}$$

Finally, we take the minimum of the RHS over all $x^\star \in X^\star$ to get the desired result. □

Theorem 14. *Assume that:*

- *the random functions f_k have a common closed domain and associated restricted strong convexity matrices M_k,*
- $\lambda_{\min}(\mathbf{E}M) > 0$,

- $\sigma = \mathbf{E}\|g(x^\star)\| < \infty$.

Using constant step size $\alpha_k = \alpha$, we get

$$\mathbf{E}\|x_k - x^\star\| \leq \gamma^{k/2}(\alpha)\|x_0 - x^\star\| + \frac{\alpha\sigma}{1-\gamma^{1/2}(\alpha)}.$$

Proof. By corollary 1, lemma 9, and a simple recursive argument, we get

$$\begin{aligned}
\mathbf{E}\|x_k - x^\star\| &\leq \mathbf{E}\|T_{\alpha f}x_{k-1} - T_{\alpha f}x^\star\| + \mathbf{E}\|T_{\alpha f}x^\star - x^\star\| \\
&\leq \gamma^{1/2}(\alpha)\mathbf{E}\|x_{k-1} - x^\star\| + \alpha\sigma \\
\cdots &\leq \gamma^{k/2}(\alpha)\|x_0 - x^\star\| + \alpha\sigma\sum_{i=0}^{k-1}\gamma^{i/2}(\alpha) \\
&\leq \gamma^{k/2}(\alpha)\|x_0 - x^\star\| + \frac{\alpha\sigma}{1-\gamma^{1/2}(\alpha)}.
\end{aligned}$$

□

Lemma 11. *Assume the same assumptions as theorem 14 but use step size $\alpha_k = C/k$. Then we have*

$$\limsup_{k\to\infty} \mathbf{E}\|x_k - x^\star\| \leq \frac{2\sigma}{\lambda_{\min}(\mathbf{E}M)}.$$

Proof. By corollary 1, lemma 9, and a simple recursive argument, we get

$$\begin{aligned}
\mathbf{E}\|x_k - x^\star\| &\leq \mathbf{E}\|T_{\alpha f}x_{k-1} - T_{\alpha f}x^\star\| + \mathbf{E}\|T_{\alpha f}x^\star - x^\star\| \\
&\leq \gamma^{1/2}(C/k)\mathbf{E}\|x_{k-1} - x^\star\| + \sigma\frac{C}{k} \\
\cdots &\leq \|x_0 - x^\star\|\prod_{i=1}^{k}\gamma^{1/2}(C/i) + \sigma C\sum_{i=1}^{k}\frac{1}{i}\prod_{j=i+1}^{k}\gamma^{1/2}(C/j) \\
&\lesssim \|x_0 - x^\star\|\exp\left(-\frac{C\lambda_{\min}(\mathbf{E}M)}{2}\sum_{i=1}^{k}1/i\right) + \sigma C\sum_{i=1}^{k}\frac{1}{i}\exp\left(-\frac{C\lambda_{\min}(\mathbf{E}M)}{2}\sum_{j=i+1}^{k}1/j\right) \\
&\sim \|x_0 - x^\star\|\exp\left(-\frac{C\lambda_{\min}(\mathbf{E}M)}{2}\log k\right) + \sigma C\sum_{i=1}^{k}\frac{1}{i}\exp\left(-\frac{C\lambda_{\min}(\mathbf{E}M)}{2}\log(k/i)\right) \\
&= \frac{\|x_0 - x^\star\|}{k^{C\lambda_{\min}(\mathbf{E}M)/2}} + \frac{\sigma C}{k^{C\lambda_{\min}(\mathbf{E}M)/2}}\sum_{i=1}^{k}i^{C\lambda_{\min}(\mathbf{E}M)/2-1} \\
&\lesssim \frac{\|x_0 - x^\star\|}{k^{C\lambda_{\min}(\mathbf{E}M)/2}} + \frac{\sigma}{\lambda_{\min}(\mathbf{E}M)/2} \to \frac{2\sigma}{\lambda_{\min}(\mathbf{E}M)}.
\end{aligned}$$

□

Theorem 15. *Assume that:*

- *the random functions f_k are almost surely twice continuously differentiable at any given point, have a common closed domain, and associated restricted strong convexity matrices M_k,*
- $\mathbf{E}M$ *is well defined,*
- $\lambda_{\min}(\mathbf{E}M) > 0$,
- $\|\nabla f(x) - \nabla f(y)\| \leq \widetilde{L}\|x - y\|$ *for all* $x, y \in \mathbf{dom} f$, *where* $\mathbf{E}\widetilde{L}^2 < \infty$,
- $\sigma = \mathbf{E}\|g(x^\star)\|^2 < \infty$,
- $x^\star$ *is in the relative interior of the domain.*

Using step sizes $\alpha_k = C/k$, we get

$$\mathbf{E}\|x_k - x^\star\|^2 = \begin{cases} \mathcal{O}(1/k) & \text{if } C\lambda_{\min}(\mathbf{E}M) > 1 \\ \mathcal{O}(\log k/k) & \text{if } C\lambda_{\min}(\mathbf{E}M) = 1 \\ \mathcal{O}(1/k^{C\lambda_{\min}(\mathbf{E}M)}) & \text{if } C\lambda_{\min}(\mathbf{E}M) < 1. \end{cases}$$

Proof. We decompose mentioned in the main article to get

$$\begin{aligned} &\mathbf{E}[\|x_k - x^\star\|^2 | x_{k-1}] = \mathbf{E}[\|T_{\alpha_k f_k} x_{k-1} - T_{\alpha_k f_k} x^\star + T_{\alpha_k f_k} x^\star - x^\star\|^2 | x_{k-1}] \\ &= \mathbf{E}\left[\|T_{\alpha_k f_k} x_{k-1} - T_{\alpha_k f_k} x^\star\|^2 + \langle T_{\alpha_k f_k} x_{k-1} - T_{\alpha_k f_k} x^\star, T_{\alpha_k f_k} x^\star - x^\star\rangle + \|T_{\alpha_k f_k} x^\star - x^\star\|^2 \big| x_{k-1}\right] \\ &\leq \gamma(\alpha_k)\|x_{k-1} - x^\star\|^2 + \alpha_k^2\sigma^2 + \mathbf{E}[\langle T_{\alpha_k f_k} x_{k-1} - T_{\alpha_k f_k} x^\star, T_{\alpha_k f_k} x^\star - x^\star\rangle | x_{k-1}]. \end{aligned}$$

The the first term is bounded by corollary 1 and the second term by lemma 9. Now we further decompose the third term to get

$$\begin{aligned} \langle T_{\alpha_k f_k} x_{k-1} - T_{\alpha_k f_k} x^\star, T_{\alpha_k f_k} x^\star - x^\star\rangle & \\ = -\alpha_k \langle x_{k-1} - x^\star, \nabla f_k(x^\star)\rangle & \quad (15) \\ + \langle x_{k-1} - x^\star, T_{\alpha_k f_k} x^\star - x^\star + \alpha_k \nabla f_k(x^\star)\rangle & \quad (16) \\ + \langle T_{\alpha_k f_k} x_{k-1} - x_{k-1}, T_{\alpha_k f_k} x^\star - x^\star\rangle & \quad (17) \\ - \langle T_{\alpha_k f_k} x^\star - x^\star, T_{\alpha_k f_k} x^\star - x^\star\rangle. & \quad (18) \end{aligned}$$

Since $x^\star \in \mathbf{ri}\,(\mathbf{dom} f)$ by assumption, we have $\mathbf{E}\nabla f(x^\star) = 0$ by lemma 6. Therefore the expectation of term (15) is 0.

Next we bound second part of term (16) to get

$$\|T_{\alpha f}x^\star - x^\star + \alpha_k \nabla f(x^\star)\| \leq \alpha \|\nabla f(x^\star - \alpha \nabla f(x^\star)) - \nabla f(x^\star)\| \leq \alpha^2 \|\nabla f(x^\star)\| \widetilde{L},$$

where the first inequality is by lemma 10 and the second by assumption. Then we take the expectation and apply Cauchy-Schwarz to get

$$\mathbf{E}\|T_{\alpha f}x^\star - x^\star + \alpha_k \nabla f(x^\star)\| \leq \alpha^2 \sigma L.$$

Finally, by applying Cauchy-Schwartz again, we get

$$\mathbf{E}(16) \leq \alpha^2 \sigma L \|x_{k-1} - x^\star\|.$$

Finally we bound (17) and (18). By Cauchy-Schwarz applied twice,

$$\mathbf{E}[\langle T_{\alpha_k f_k} x_{k-1} - x_{k-1}, T_{\alpha_k f_k} x^\star - x^\star \rangle | x_{k-1}] \leq \alpha_k^2 \sqrt{\mathbf{E}[\|\nabla f_k(x_{k-1})\|^2 | x_k] \mathbf{E}[\|\nabla f_k(x^\star)\|^2]}.$$

Note that the assumption

$$\|\nabla f(x) - \nabla f(x^\star)\| \leq \widetilde{L} \|x - x^\star\|$$

implies

$$\|\nabla f(x)\| \leq \widetilde{L} \|x - x^\star\| + \|\nabla f(x^\star)\|$$

by the triangle inequality. By applying this, we get

$$\mathbf{E}\left[\langle T_{\alpha_k f_k} x_{k-1} - x_{k-1}, T_{\alpha_k f_k} x^\star - x^\star \rangle | x_{k-1}\right] \leq \alpha_k^2 (L \|x_{k-1} - x^\star\| + \sigma)\sigma.$$

So finally, we combine all these results and lemma 11 and get

$$
\begin{aligned}
\mathbf{E}\|x_k - x^\star\|^2 &\leq \gamma(\alpha_k)\|x_{k-1} - x^\star\|^2 + \alpha_k^2\left(3\sigma^2 + 2\sigma L\|x - x^\star\|\right) \\
&\sim \prod_{i=1}^{k}\gamma(C/i)\|x_0 - x^\star\|^2 + \sum_{i=1}^{k}\frac{C^2}{i^2}(3\sigma^2 + 4\sigma^2 L/\lambda_{\min}(\mathbf{E}M))\prod_{j=i+1}^{k}\gamma(C/i) \\
&\lesssim \|x_0 - x^\star\|^2 \exp\left(-C\lambda_{\min}(\mathbf{E}M)\sum_{i=1}^{k}1/i\right) \\
&\qquad + \sum_{i=1}^{k}\frac{C^2}{i^2}(3\sigma^2 + 4\sigma^2 L/\lambda_{\min}(\mathbf{E}M))\exp\left(-C\lambda_{\min}(\mathbf{E}M)\sum_{j=i+1}^{k}1/i\right) \\
&\sim \frac{\|x_0 - x^\star\|^2}{k^{C\lambda_{\min}(\mathbf{E}M)}} + \frac{C^2(3\sigma^2 + 4\sigma^2 L/\lambda_{\min}(\mathbf{E}M))}{k^{C\lambda_{\min}(\mathbf{E}M)}}\sum_{i=1}^{k} i^{C\lambda_{\min}(\mathbf{E}M)-2} \\
&\sim \frac{\|x_0 - x^\star\|^2}{k^{C\lambda_{\min}(\mathbf{E}M)}} + C^2\left(3\sigma^2 + \frac{4\sigma^2 L}{\lambda_{\min}(\mathbf{E}M)}\right)\begin{cases} \frac{1}{k(C\lambda_{\min}(\mathbf{E}M)-1)} & \text{if } C\lambda_{\min}(\mathbf{E}M) > 1 \\ \frac{\log k}{k} & \text{if } C\lambda_{\min}(\mathbf{E}M) = 1 \\ \frac{1}{k^{C\lambda_{\min}(\mathbf{E}M)}(1 - C\lambda_{\min}(\mathbf{E}M))} & \text{if } C\lambda_{\min}(\mathbf{E}M) < 1 \end{cases} \\
&\sim \begin{cases} \frac{C^2(3\sigma^2 + 4\sigma^2 L/\lambda_{\min}(\mathbf{E}M))}{(C\lambda_{\min}(\mathbf{E}M)-1)}\frac{1}{k} & \text{if } C\lambda_{\min}(\mathbf{E}M) > 1 \\ C^2(3\sigma^2 + 4\sigma^2 L/\lambda_{\min}(\mathbf{E}M))\frac{\log k}{k} & \text{if } C\lambda_{\min}(\mathbf{E}M) = 1 \\ \left(\frac{1}{(1 - C\lambda_{\min}(\mathbf{E}M))} + \|x_0 - x^\star\|^2\right)\frac{1}{k^{C\lambda_{\min}(\mathbf{E}M)}} & \text{if } C\lambda_{\min}(\mathbf{E}M) < 1. \end{cases}
\end{aligned}
$$

Finally, we get the desired result by disregarding all the constants. □

E Counter examples

Example 1. *Non-uniqueness of the "best" restricted strong convexity matrix.*

Consider the convex function

$$f(x, y) = \max\{x^2 + 2y^2, 2x^2 + y^2\}$$

which has restricted strong convexity matrices

$$M_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad M_2 = \begin{bmatrix} 2/3 & 2/3 \\ 2/3 & 2/3 \end{bmatrix}.$$

We can verify via brute force that M_1 and M_2 are *maximal*. More precisely, f has no restricted strong convexity matrix M such that $M \neq M_1$ and $M \succeq M_1$, and the same holds for M_2. In particular, M_1 and M_2 are *incomparable*, *i.e.*, $M_1 \not\succeq M_2$ and $M_1 \not\preceq M_2$.

So given a convex function, it may not always be possible to uniquely choose the "strongest" restricted strong convexity matrix. As this example shows, it is possible to have many distinct choices that are incomparable.

Fundamentally, we have this complication because the set of symmetric matrices (which is partially ordered) does not have a well-defined *meet*, *i.e.*, the partially ordered set does not form a lattice. This example was inspired by [Mak].

Example 2. *$T_{\breve{f}+r}$ is not always a stronger contraction than $T_{\breve{f}}$.*

Consider the functions

$$\breve{f}(x) = \frac{1}{2} x^T \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} x \qquad r(x) = \frac{1}{2} x^T \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} x$$

and the points

$$x = \begin{bmatrix} 3 \\ 0 \end{bmatrix} \qquad y = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Then simple calculations gives us

$$\|x - y\|^2 = 10 \qquad \|T_{\breve{f}} x - T_{\breve{f}} y\|^2 = 3.25 \qquad \|T_{\breve{f}+r} x - T_{\breve{f}+r} y\|^2 = 3.4$$

and we see that $T_{\breve{f}+r}$ is a *weaker* contraction than $T_{\breve{f}}$. However, theorem 10 is not violated since $3.4/10 \leq \sqrt{3.25/10}$.

Fundamentally, we have this complication because $A \succeq B \succeq 0$ does not imply $A^2 \succeq B^2$. This example was inspired by [CK85].

References

[BC11] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2011.

[Ber99] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 2nd edition, 1999.

[Ber11] D. P. Bertsekas. Incremental proximal methods for large scale convex optimization. *Mathematical Programming*, 129(2):163–195, 2011.

[Bil95] P. Billingsley. *Probability and Measure*. Wiley-Interscience, 3rd edition, 1995.

[BM11] F. Bach and E. Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. C. N. Pereiraand, and K Q. Weinberger, editors, *NIPS*, pages 451–459, 2011.

[BMMW12] H. H. Bauschke, S. M., Moffat, and X. Wang. Firmly nonexpansive mappings and maximally monotone operators: Correspondence and duality. *Set-Valued and Variational Analysis*, 20(1):131–153, 2012.

[BV04] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[CK85] N. N. Chan and M. K. Kwong. Hermitian matrix inequalities and a conjecture. *The American Mathematical Monthly*, 92(8):533–541, 1985.

[DS09] J. C. Duchi and Y. Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10:2899–2934, 2009.

[GL12] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, 4th edition, 2012.

[HJ90] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1990.

[KB10] B. Kulis and P. L. Bartlett. Implicit online learning. In J. Fürnkranz and T. Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning (ICML2010)*, pages 575–582, Haifa, Israel, 2010. Omnipress.

[KL11] N. Karampatziakis and J. Langford. Online importance weight aware updates. In F. G. Cozman and A. Pfeffer, editors, *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, pages 392–399, Barcelona, Spain, 2011. AUAI Press.

[KY03] H. J. Kushner and G. G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2nd edition, 2003.

[LLZ09] J. Langford, L. Li, and T. Zhang. Sparse online learning via truncated gradient. *Journal of Machine Learning Research*, 10:777–801, 2009.

[Mak] Y. Makarychev. Properties of the cone of positive semidefinite matrices. Mathematics Stack Exchange. URL:http://math.stackexchange.com/q/395782 (version: 2013-05-19).

[McM11] H. B. McMahan. Follow-the-regularized-leader and mirror descent: Equivalence theorems and l1 regularization. In G. Gordon, D. Dunson, and M. Dudk, editors, *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*. MIT Press, 2011.

[Mor65] J. J. Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Socété Mathématique de France*, 93:273–299, 1965.

[NJLS09] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[NY78] A. S. Nemirovski and D. B. Yudin. Cesari convergence of the gradient method of approximating saddle points of convex-concave functions (in Russian). *Doklady Akademii Nauk SSSR*,

239(5), 1978. Soviet Mathematics Doklady, 19(2), 1978 (in English).

[NY83] A. S. Nemirovski and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization.* Wiley, 1983.

[PB13] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):123–231, 2013.

[PJ92] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, July 1992.

[Pol87] B. T. Polyak. *Introduction to Optimization*. Optimization Software, Inc., 1987.

[RM51] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

[Roc70] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[Roc76] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

[Rup88] D. Ruppert. Efficient estimations from a slowly convergent robbins-monro process. Technical report, Operations Research and Industrial Engineering, Cornell University, Ithaca, NY, 1988.

[SD03] R. J. Santos and Á. R. De Pierro. A cheaper way to compute generalized cross-validation as a stopping rule for linear stationary iterative methods. *Journal of Computational and Graphical Statistics*, 12(2):417–433, 2003.

[Sho62] N. Z. Shor. *Notes of Scientic Seminar on Theory and Applications of Cybernetics and Operations Research*, page 917, 1962.

[SKR85] N. Z. Shor, K. C. Kiwiel, and A. Ruszcayǹski. *Minimization Methods for Non-differentiable Functions*. Springer-Verlag New York, Inc., 1985.