Projected Gradient Methods

Benjamin Recht

Department of Computer Sciences, University of Wisconsin-Madison
1210 W Dayton St, Madison, WI 53706
email: brecht@cs.wisc.edu

November 9, 2012

1 Proximal Point Mappings Associated with Convex Functions

Let P be an extended-real-valued convex function on $\mathbb{R}^n$. Define the operator

$$\mathrm{prox}_P(x) = \arg\min_y \tfrac{1}{2}\|x-y\|_2^2 + P(y) \tag{1.1}$$

Since the optimized function is strongly convex, it must have a unique optimal solution. Therefore, we can conclude that $\mathrm{prox}_P(x)$ is a well-defined mapping from $\mathbb{R}^n$ to $\mathbb{R}^n$. By the first order optimality conditions, we conclude that $\mathrm{prox}_P(x)$ is the unique point satisfying

$$x - \mathrm{prox}_P(x) \in \partial P(\mathrm{prox}_P(x))\,. \tag{1.2}$$

The definition of prox_P also reveals that it is well-defined for all $x \in \mathbb{R}^n$, and maps onto the set $\mathrm{dom}(P) := \{z \in \mathbb{R}^n \;:\; P(z) < \infty\}$. The mapping prox_P is called the *proximity operator* or *proximal point mapping* associated with P.

Let's look at some examples.

1. If $\mathbb{I}_C$ is an indicator function for a convex set C

$$\mathbb{I}_C(x) = \begin{cases} 0 & x \in C \\ \infty & \text{otherwise} \end{cases} \tag{1.3}$$

   then $\mathrm{prox}_{\mathbb{I}_C}$ is the Euclidean projection onto C. That is, $\mathrm{prox}_{\mathbb{I}_C}(x)$ is the closest point in the set C to x in Euclidean distance.

2. For $\mathbb{I}_{\mathbb{R}_+}$, this proximity mapping takes on the trivial form:

$$\mathbb{I}_{\mathbb{R}_+}(x)_i = \max(x_i, 0) \tag{1.4}$$

3. For $P(x) = \frac{\mu}{2}\|x\|_2^2$, $\mathrm{prox}_P(x) = \frac{1}{1+\mu}x$. That is, $\mathrm{prox}_P(x)$ is equal to a multiple of x, shrunk towards the origin.

4. For $P(x) = \mu\|x\|_1$,

$$\operatorname{prox}_P(x)_i = \begin{cases} x_i + \mu & x_i < -\mu \\ 0 & -\mu \leq x_i \leq \mu \\ x_i - \mu & x_i > \mu \end{cases} \tag{1.5}$$

   This function is called the *shrinkage* operator and has many applications in signal processing. To see that this is the correct form, one needs only to analyze the optimality conditions of the one dimensional problem

$$\text{minimize } \tfrac{1}{2}(x-y)^2 + \mu|y| \tag{1.6}$$

2 The Proximal Point Algorithm

Proximity operators have many algorithmic applications. As a warm up, consider the following simple iteration: pick $x_0 \in \mathbb{R}^n$ and $\nu > 0$ and define the iteration $x_{k+1} = \operatorname{prox}_{\nu P}(x_k)$. This simple iteration can be shown to converge to a minimizer of the function P. To prove this, we need the following two lemmas. The first is a simple consequence of the convexity of P.

Lemma 2.1 *Let P be convex on X. Let x, $y \in X$, and let $g_y \in \partial P(y)$ and $g_x \in \partial P(y)$. Then* $\langle g_x - g_y, x - y\rangle \geq 0$.

Proof By the definition of the subdifferential, we have

$$\begin{aligned} P(x) - P(y) &\geq \langle g_y, x - y\rangle \\ P(y) - P(x) &\geq \langle g_x, y - x\rangle \end{aligned} \tag{2.1}$$

Adding these two equations gives $-\langle g_x - g_y, x - y\rangle \leq 0$. ∎

The second lemma uses this key inequality to establish several facts about the proximity operator. This lemma is proven in [1].

Lemma 2.2 *Let $Q_\nu(x) := x - \operatorname{prox}_{\nu P}(x)$. Then we have*

(i) $\nu^{-1}Q_\nu(x) \in \partial P(\operatorname{prox}_{\nu P}(x))$

(ii) $\langle \operatorname{prox}_{\nu P}(x) - \operatorname{prox}_{\nu P}(z), Q_\nu(x) - Q_\nu(z)\rangle \geq 0$

(iii) $\|\operatorname{prox}_{\nu P}(x) - \operatorname{prox}_{\nu P}(z)\|^2 + \|Q_\nu(x) - Q_\nu(z)\|^2 \leq \|x - z\|^2$

(iv) $\|x - z\| = \|\operatorname{prox}_{\nu P}(x) - \operatorname{prox}_{\nu P}(z)\|$ *if and only if* $x - z = \operatorname{prox}_{\nu P}(x) - \operatorname{prox}\, \nu P(z)$

Proof The first assertion follows from the definitions. The second assertion follows from (i), and Lemma 2.1. The third assertion follows from (ii) after expanding the identity

$$\|x - z\|^2 = \|[P_\nu(x) - P_\nu(z)] + [Q_\nu(x) - Q_\nu(z)]\|^2 .$$

(iv) follows immediately from (iii). ∎

By Lemma 2.2 (iii), we have

$$\|\operatorname{prox}_{\nu P}(x) - \operatorname{prox}_{\nu P}(z)\|^2 \leq \|x - z\|^2 \tag{2.2}$$

and we say that the proximity operator is *nonexpansive*. This is the essential property needed to prove the convergence of the proximal point method. That the proximity operator is nonexpansive also plays a role in the projected gradient algorithm, analyzed below.

Using the nonexpansive property of the proximity operator, we can now verify the convergence of the proximal point method. Since $\operatorname{prox}_{\nu P}$ is non-expansive, $\{z_k\}$ lies in a compact set and must have a limit point $\bar{z}$. Also for any z_* with $0 \in \partial P(z_*)$,

$$\|z_{k+1} - z_*\| = \|\operatorname{prox}_{\nu P}(z_k) - \operatorname{prox}_{\nu P}(z_*)\| \leq \|z_k - z_*\| \tag{2.3}$$

which means that the sequence $\|z_k - z_*\|$ is monotonically non-increasing. Therefore

$$\lim_{k\to\infty} \|z_k - z_*\| = \|\bar{z} - z_*\| \,. \tag{2.4}$$

where $\bar{z}$ is any limit point of z_k. By continuity we have $\operatorname{prox}_{\nu P}(\bar{z})$ is also a limit point of z_k. Therefore, we must have

$$\|\operatorname{prox}_{\nu P}(\bar{z}) - \operatorname{prox}_{\nu P}(z_*)\| = \|\operatorname{prox}_{\nu P}(\bar{z}) - z_*\| = \|\bar{z} - z_*\| \tag{2.5}$$

But this means that $\operatorname{prox}_{\nu P}(\bar{z}) - \operatorname{prox}_{\nu P}(z_*) = \bar{z} - z_*$, and in turn that $\operatorname{prox}_{\nu P}(\bar{z}) = \bar{z}$ and $0 \in \partial P(\bar{z})$. Now using $\bar{z}$ for z_* in (2.4) shows that

$$\lim_{k\to\infty} \|z_k - \bar{z}\| = 0 \tag{2.6}$$

In other words, the sequence z_k converges to $\bar{z}$.

3 The projected gradient algorithm

The projected gradient algorithm combines a proximal step with a gradient step. This lets us solve a variety of constrained optimization problems with simple constraints, and it lets us solve some non-smooth problems at linear rates.

We will aim to analyze a function h which admits a decomposition

$$h(x) = f(x) + P(x) \tag{3.1}$$

where f is smooth and P is a convex extended real valued function. Let us assume that ∇f is Lipschitz so that $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$.

Let us define a projected gradient scheme to solve this problem. Let $\alpha_0, \ldots, \alpha_T, \ldots,$ be a sequence of positive step sizes. Choose $x_0 \in X$, and iterate

$$x_{k+1} = \operatorname{prox}_{\alpha_k P}(x_k - \alpha_k \nabla f(x_k)) \,. \tag{3.2}$$

The algorithm alternates between taking gradient steps and then taking proximal point steps.

The key idea behind this algorithm is summed up by the following proposition

Proposition 3.1 *Let f be differentiable and convex and let P be convex. x_* is an optimal solution of*

$$\textit{minimize}_x \; f(x) + P(x) \tag{3.3}$$

if and only if $x_* = \operatorname{prox}_{\nu P}(x_* - \nu \nabla f(x_*))$ for all $\nu > 0$.

Proof x_* is an optimal solution if and only if $-\nabla f(x_*) \in \partial P(x_*)$. This is equivalent to

$$(x_* - \nu \nabla f(x_*)) - x_* \in \nu \partial P(x_*)\,,$$

which is equivalent to $x_* = \mathrm{prox}_{\nu P}(x_* - \nu \nabla f(x_*))$. ■

For non-convex f, we see that a fixed point of the projected gradient iteration is a stationary point of h. We first analyze the convergence of this projected gradient method for arbitrary smooth f, and then focus on strongly convex f.

3.1 General Case

Let h_* denote the optimal value of (3.1). Suppose we set $\alpha_k = 1/M$ for all k with $M \geq L$. Then we have

$$\|x_{k+1} - x_k\| \leq \sqrt{\frac{2(h(x_0) - h_*)}{M(k+1)}}\,. \tag{3.4}$$

This expression confirms that x_k will converge to some fixed point.

To verify this inequality, note that for any x, y,

$$h(x) = f(x) + P(x) \leq f(y) + \nabla f(y)^T(x-y) + \frac{M}{2}\|x-y\|^2 + P(x) =: u(x;y) \tag{3.5}$$

for any $M \geq L$. This is just Taylor's series. Note that the minimizer of $u(x;y)$ (with respect to x) is equal to

$$\mathrm{prox}_{P/M}(y - 1/M \nabla f(y))\,. \tag{3.6}$$

and also note that $u(x;y)$ is strongly convex with parameter M.

Now we have the chain of inequalities

$$h(x_k) - h(x_{k+1}) \geq h(x_k) - u(x_{k+1}; x_k) \tag{3.7}$$

$$= u(x_k; x_k) - u(x_{k+1}; x_k) \tag{3.8}$$

$$\geq \frac{M}{2}\|x_{k+1} - x_k\|^2 \tag{3.9}$$

Summing these inequalities up for $k = 1, \ldots, n$, we have

$$\sum_{k=0}^{n} \|x_{k+1} - x_k\|^2 \leq \frac{2}{M}(h(x_0) - h_*) \tag{3.10}$$

and the conclusion follows

3.2 Strongly Convex Case

Let's now assume that f is strongly convex with strong convexity parameter ℓ:

$$f(z) \geq f(x) + \nabla f(x)^*(z - x) + \frac{\ell}{2}\|z - x\|^2\,. \tag{3.11}$$

Let x_* denote the optimal solution of (3.1). x_* is unique because of strong convexity. Observe that

$$\|x_{k+1} - x_*\| = \| \operatorname{prox}_{\alpha_k P}(x_k - \alpha_k \nabla f(x_k) - \operatorname{prox}_{\alpha_k P}(x_* - \alpha_k \nabla f(x_*))\| \tag{3.12}$$
$$\leq \|x_k - \alpha_k \nabla f(x_k) - x_* + \alpha_k \nabla f(x_*)\|] \tag{3.13}$$

Here, the first equality follows by the definition of x_{k+1} and because x_* is optimal (see Proposition 3.1). (3.13) follows from Proposition 2.2.

Since f is strongly convex and has a Lipschitz continuous gradient, it follows that for all vectors x and y and all positive scalars t

$$\|x - \nu \nabla f(x) - (y - \nu \nabla f(y))\| \leq \max\{|1 - \nu L|, |1 - \nu \ell|\} \|x - y\| \,. \tag{3.14}$$

To see this, note that

$$\|x - t\nabla f(x) - (y - t\nabla f(y))\| \leq \left\| \int_0^1 (I - t\nabla^2 f(x + t(y-x))(y-x)dt \right\| \tag{3.15}$$
$$\leq \sup_z \|I - t\nabla^2 f(z)\| \|y - z\| \,. \tag{3.16}$$

Note that the minimum eigenvalue of $\nabla^2 f(z)$ is at least ℓ and the maximum eigenvalue is at least L. Therefore the eigenvalues of $I - t\nabla^2 f(z)$ are at most $\max(1 - tL, 1 - t\ell)$ and at least $\min(1 - tL, 1 - t\ell)$. Therefore, $\|I - t\nabla^2 f(z)\| \leq \max(|1 - tL|, |1 - t\ell|)$.

In particular, using this upper bound in (3.13), we have

$$\|x_{k+1} - x_*\| \leq \max\{|1 - \alpha_k L|, |1 - \alpha_k \ell|\} \|x - y\| \,. \tag{3.17}$$

Note that $\alpha_k = \frac{2}{L+\ell}$ minimizes the right hand side for all k. Setting α_k to this value, we find that

$$\|x_{k+1} - x_*\| \leq \left(\frac{L - \ell}{L + \ell} \right) \|x_k - x_*\| \tag{3.18}$$

or, denoting $\kappa = \frac{L}{\ell}$ and $D_0 = \|x_0 - x_*\|$,

$$\|x_k - x_*\| \leq \left(\frac{\kappa - 1}{\kappa + 1} \right)^k D_0 \tag{3.19}$$

That is, for strongly convex f and arbitrary P, the projected gradient algorithm converges at a linear rate under a constant step-size policy.

4 Constrained Optimization

Let C be a convex set and let $\mathbb{I}_C$ denote its indicator function. What's the subdifferential of $\mathbb{I}_C(x)$ for $x \in C$? By definition $g \in \partial \mathbb{I}_C(x)$ if and only if

$$\mathbb{I}_C(y) \geq \mathbb{I}_C(x) + g^T(y - x) \tag{4.1}$$

for all y. This is equivalent to

$$\partial \mathbb{I}_C(x) = \{g \,:\, g^T(x - y) \geq 0 \; \forall y \in C\} \tag{4.2}$$

for $x \in C$. This set is often called the *normal cone* of C at x.

Consider the constrained optimization problem

$$\text{minimize}_{x \in C} f(x) \tag{4.3}$$

for smooth, convex f. Then x_* is optimal if and only if $-\nabla f(x_*) \in \partial \mathbb{I}_C(x_*)$. We can find such an x_* via the projected gradient algorithm.

References

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